A multicentre study of validity and reliability of measures of cold challenge measured by LSCI and thermography: outcome measures for SSc-related RP

Journal:	*Arthritis & Rheumatology*
Manuscript ID	ar-17-1194.R2
Wiley - Manuscript type:	Full Length
Date Submitted by the Author:	n/a
Complete List of Authors:	Wilkinson, Jack; University of Manchester, Salford Royal Research & Development and Centre for Biostatistics Leggett, Sarah; The University of Manchester, School of Biological Sciences, Faculty of Biology, Medicine and Health, Division of Musculoskeletal and Dermatological Sciences, University of Manchester Manchester Academic Health Science Centre, Salford Royal NHS Foundation, Marjanovic, Elizabeth; University of Manchester, School of Biological Sciences, Faculty of Biology, Medicine and Health, Division of Musculoskeletal and Dermatological Sciences, University of Manchester Manchester Academic Health Science Centre, Salford Royal NHS Foundation, Moore, Tonia; University of Manchester, School of Biological Sciences, Faculty of Biology, Medicine and Health, Division of Musculoskeletal and Dermatological Sciences, University of Manchester Manchester Academic Health Science Centre, Salford Royal NHS Foundation, Allen, John; Freeman Hospital, Microvascular Diagnostics, Northern Medical Physics and Clinical Engineering; Newcastle University, Institute of Cellular Medicine, Medical School Anderson, Marina; University of Liverpool, Institute of Clinical Sciences Britton, Jason; Leeds Teaching Hospitals NHS Trust, Medical Physics Department Buch, Maya; University of Leeds, Leeds Institute of Rheumatic & Musculoskeletal Medicine; The National Institute for Health Research (NIHR) Leeds Musculoskeletal Biomedical Research Unit Del Galdo, Francesco; University of Leeds, Leeds Institute of Rheumatic and Musculoskeletal Medicine; The National Institute for Health Research (NIHR) Leeds Musculoskeletal Biomedical Research Unit Denton, Christopher; Royal Free Hospital and UCL Medical School, Centre for Rheumatology Dinsdale, Graham; The University of Manchester, Centre for Musculoskeletal Research, School of Biological Sciences, Faculty of Biology, Medicine and Health, Division of Musculoskeletal and Dermatological Sciences, University of Manchester Manchester Academic Health Science Centre, Salford Royal NHS Foundation

	Griffiths, Bridget; Freeman Hospital, Department of Rheumatology Hall, Frances; Cambridge University Hospitals NHS Foundation Trust, Department of Rheumatology Howell, Kevin; University College London Medical School, Centre for Rheumatology and Connective Tissue Diseases MacDonald, Audrey ; Freeman Hospital, Microvascular Diagnostics, Northern Medical Physics and Clinical Engineering; Newcastle University, Institute of Cellular Medicine, Medical School McHugh, Neil; Royal United Hospitals NHS Foundation Trust, Department of Rheumatology, Royal National Hospital for Rheumatic Diseases Manning, Joanne; Salford Royal NHS Foundation Trust, Department of Rheumatology Pauling, John; University of Bath, Department of Pharmacy and Pharmacology; Royal National Hospital for Rheumatic Diseases, Rheumatology Roberts, Christopher; University of Manchester, Biostatistics Department Shipley, Jacqueline; Royal United Hospitals NHS Foundation Trust, Department of Rheumatology, Royal National Hospital for Rheumatic Diseases Herrick, Ariane; University of Manchester, School of Biological Sciences, Faculty of Biology, Medicine and Health, Division of Musculoskeletal and Dermatological Sciences, University of Manchester Manchester Academic Health Science Centre, Salford Royal NHS Foundation, Murray, Andrea; University of Manchester, School of Biological Sciences, Faculty of Biology, Medicine and Health, Division of Musculoskeletal and Dermatological Sciences, University of Manchester Manchester Academic Health Science Centre, Salford Royal NHS Foundation, ; University of Manchester, Photon Science Institute
Keywords:	Scleroderma, Raynaud’s Phenomenon, Imaging, Outcome Measures, Patient Reported Outcomes
<B>Disease Category: Please select the category from the list below that best describes the content of your manuscript.:	Systemic Sclerosis

Running head: Multicentre validity and reliability of hand cold challenge as measured by LSCI and thermography

Title: A multicentre study of validity and reliability of responses to hand cold challenge as measured by laser speckle contrast imaging and thermography: outcome measures for systemic sclerosis-related Raynaud's phenomenon

Jack D Wilkinson[1], BSc, MSc

Sarah A Leggett[2], BSc

Elizabeth J Marjanovic[2], BSc, PhD

Tonia L Moore[2], BSc

John Allen[3,4], BSc, PhD

Marina E Anderson[5], FRCP PhD

Jason Britton[6], BSc MSc

Maya H Buch7[7,8], MBChB, FRCP, PhD

Francesco Del Galdo[7,8],MD, PhD

Christopher P Denton[9], PhD, FRCP

Graham Dinsdale[2], MPhys, PhD

Bridgett Griffiths[10], MB ChB, MD, FRCP

Frances Hall[11], MAoxon, FRCP, DPhil

Kevin Howell[9], BSc, MSc, PhD

Audrey MacDonald[3,4], BSc MSc

Neil J McHugh[12], MBChB, MD

Joanne B Manning[13], HNC

John D Pauling[12,14] , BMedSci, PhD, FRCP

Christopher Roberts[1], BSc,PhD

Jacqueline A Shipley[12], BSc, MSc, PhD

Ariane L Herrick[2,15], MD, FRCP

Andrea K Murray[2,16], MPhys, PhD

Corresponding author:

Andrea Murray

Rm C214, Clinical Sciences Building,

Centre for Musculoskeletal Research,

School of Biological Sciences,

Faculty of Biology, Medicine and Health,

Division of Musculoskeletal and Dermatological Sciences,

University of Manchester,

Manchester Academic Health Science Centre,

Salford Royal NHS Foundation,

Manchester,

M13 9PT,

UK.

Andrea.murray@manchester.ac.uk

Tel +44 (0)161 206 1538/1133

Addresses:

1. Centre for Biostatistics,

University of Manchester,

Manchester,

UK

2. Centre for Musculoskeletal Research,

School of Biological Sciences,

Faculty of Biology, Medicine and Health,

Division of Musculoskeletal and Dermatological Sciences,

University of Manchester

Manchester Academic Health Science Centre,

Salford Royal NHS Foundation,

Manchester,

UK

3. Microvascular Diagnostics,

Northern Medical Physics and Clinical Engineering,

Freeman Hospital,

Newcastle upon Tyne,

UK

4. Institute of Cellular Medicine,

Medical School,

Newcastle University,

Newcastle upon Tyne,

UK

5. Institute of Clinical Sciences,

University of Liverpool,

Liverpool,

UK

6. Medical Physics Department,

Leeds Teaching Hospitals NHS Trust,

Leeds,

UK

7. Leeds Institute of Rheumatic and Musculoskeletal Medicine,

Chapel Allerton Hospital,

Leeds,

UK

8. The National Institute for Health Research (NIHR) Leeds Musculoskeletal Biomedical Research Unit,

Chapel Allerton Hospital,

Leeds,

UK

9. Centre for Rheumatology and Connective Tissue Diseases,

UCL Medical School,

Royal Free Campus,

London,

UK

10. Department of Rheumatology,

Freeman Hospital,

Newcastle upon Tyne Hospitals NHS Foundation Trust,

Newcastle upon Tyne,

UK

11. Department of Rheumatology,

Cambridge University Hospitals NHS Foundation Trust,

Cambridge,

UK

12. Department of Rheumatology,

Royal National Hospital for Rheumatic Diseases,

Royal United Hospitals NHS Foundation Trust,

Upper Borough Walls,

Bath,

UK

13. Department of Rheumatology,

Salford Royal NHS Foundation,

Salford,

UK

14. Department of Pharmacy and Pharmacology,

University of Bath,

Claverton Downs,

Bath,

UK

15. NIHR Manchester Musculoskeletal Biomedical Research Centre,

Central Manchester NHS Foundation Trust,

Manchester Academic Health Science Centre,

Manchester

UK

16. Photon Science Institute

University of Manchester,

Manchester

UK.

FUNDING

This study was funded by an Arthritis Research UK Clinical Studies Grant [20656] and an unrestricted educational grant from Actelion Pharmaceuticals Ltd. This study was supported

by Arthritis Research UK grant number (Centre for Epidemiology) 20380 and the Manchester Academic Health Sciences Centre (MAHSC).

COMPETING INTERESTS

CP Denton - GSK, Celgene, Actelion, Bayer, Sanofi, Roche-Genentech, Inventiva, 5,CSL Behring, GSK, Actelion, Roche-Genentech, Inventiva.

JD Pauling, has received unrestricted grant support (totalling £25000), speaker honoraria and consultancy fees from Actelion Pharmaceuticals.

AL Herrick has done consultancy work for Actelion, served on a Data Safety Monitoring Board for Apricus, received research funding and speaker's honoraria from Actelion, and speaker's honoraria from GSK.

AK Murray has received speaker honoraria from GSK and speaker honoraria and the grant associated with this study (referenced above) from Actelion.

All other authors declare no conflicts or competing interests.

Background: Objective and reliable outcome measures to facilitate clinical trials of novel treatments for systemic sclerosis (SSc)-related Raynaud's phenomenon (RP) are badly needed. Laser speckle contrast imaging (LSCI) and thermography are non-invasive measures of perfusion that show excellent potential. The purpose of this multi-centre study was to determine the reliability and validity of a hand cold challenge protocol using LSCI, standard thermography and low-cost mobile phone-based thermography.

Methods: Patients with RP secondary to SSc were recruited from 6 UK tertiary-SSc centres and underwent cold challenge on 2 consecutive days. Changes in cutaneous perfusion/temperature at each visit were imaged simultaneously using LSCI, standard and mobile phone thermography. Measurements included area under reperfusion/rewarming curve (AUC) and maximum perfusion/rewarming (MAX). Test-retest reliability was assessed using intraclass correlation coefficients (ICC). Estimated latent correlations assessed convergent validity of LSCI and thermography.

Results: 159 patients were recruited (84% female, 77% limited cutaneous, median age 63.3 years). LSCI and standard thermography both had substantial reliability, ICCs (95%CI) for AUC were 0.67(0.54-0.76) and 0.68(0.58-0.80) respectively, and for MAX were 0.64(0.52-0.75) and 0.72(0.64-0.81) respectively. Very high latent correlations (95% CI) were present for AUCs of LSCI and thermography [0.94(0.87-1.00)], and for AUCs of standard and mobile phone thermography [0.98(0.94-1.00)].

Conclusion: This is the first multi-centre study examining reliability and validity of cold challenge using LSCI and thermography in patients with SSc-related RP. LSCI and thermography demonstrated good potential as outcome measures. LSCI, standard and mobile phone thermography had very high convergent validity.

Systemic sclerosis (SSc)-related digital vasculopathy is painful and disabling, and has significant impact on quality of life. Raynaud's phenomenon (RP) occurs in most patients with SSc (96%) and is consistently the highest ranked symptom of SSc in terms of frequency and impact on daily function (1,2). In patients with SSc, RP often progresses to severe digital vasculopathy, with up to 50% of patients developing painful digital ulceration (3-11). Treatments are far from ideal and Cochrane and other reviews highlight the lack of evidence base for the treatment of both primary and SSc-related RP (12-15): one of the reasons for this is the lack of reliable outcome measures that are necessary to deliver successful clinical trials. Technological advances in laboratory measurements of blood flow (laser speckle contrast imaging [LSCI] and thermography [skin temperature- a pseudo measure of perfusion]) hold promise as objective measures (16,17). The Outcome Measures in Rheumatology (OMERACT) 6 report, describing the current status of outcome measure development for clinical trials in SSc, concluded that whether imaging techniques made the transition from research pathophysiology measurement techniques to outcome measures for RP was dependent upon 'whether data are published or available to show their validity' (18). The requirement for reliable outcome measures to facilitate highly powered clinical trials in SSc-related RP is now especially pertinent due to on-going novel drug developments (19-23). Whilst patient-reported outcome measures such as the Raynaud's condition score (RCS(24)) are well suited for later, phase III studies, objective non-invasive imaging techniques would provide confirmatory testing to inform stop-go decision-making in earlier phase II studies.

Our main aim was to determine whether LSCI and thermography, alongside a cold challenge of the hands, are sufficiently reliable and valid to allow their use as outcome measures in

multicentre clinical trials. Our primary objectives were to evaluate test-retest reliability and construct validity (25), which we defined as the ability of LSCI and thermography to measure important features of SSc-related digital vasculopathy. Our secondary objectives were to assess inter-observer reliability, and feasibility, of the techniques. Just prior to commencement of our study, mobile phone thermography came on the market as an imaging method, potentially offering a more cost-effective and portable alternative to LSCI and 'standard' thermography. Hence an additional secondary objective was to assess the utility of mobile phone thermography in comparison to standard thermography.

METHODS

Six UK tertiary-SSc centres took part in the study; individuals responsible for imaging and analysis attended a central training session prior to the start of recruitment; At least one person from each centre attended the training.

Patients

The study aimed to recruit 180 patients with SSc. Inclusion and exclusion criteria are listed in supplementary Table S1 and included current digital ulceration. The study was approved by Cambridgeshire and Hertfordshire NRES committee (15/EE/0083) and all patients gave written consent. Each visit took approximately one hour. All patients were recruited between 1/10/15 and 28/2/15 to minimise inter-individual variation related to season.

Imaging equipment

An LSCI (FLPI-2 [Moor Instruments, Axminster, UK] Figures 1a and 2b (16,17)) was leased to each centre. Five of the six centres used their own thermal cameras (referred to as 'standard thermography' Figure 1a, b and d (26)); a camera was leased to the sixth. A mobile phone/device-connectable thermography camera (FLIR One Figure 1a and 1c) and an Apple iPhone 5 was purchased for each centre along with all other cold challenge equipment (to minimise centre variation). To minimise differences between centres, equipment at each site was set-up according to strict guidelines for positioning to ensure images were taken in as similar manner as possible (angles/distances) and underwent a calibration protocol at the start and end of the study (carried out by a single person from the central site).. LSCI settings were adjusted for distance, frequency, duration, focus, intensity overlay, processing mode (high resolution), and colour image acquisition. Thermal cameras settings were adjusted for room temperature, distance to hands and skin emissivity. Mobile phone thermography settings were limited but set to matt.

Cold challenge

Patients were requested to wear light clothing and refrain from vigorous exercise, caffeine and alcohol for 4 hours prior to the assessment. Upon arrival patients were seated comfortably for 20 minutes and acclimatised; clinical research forms were completed. Immediately prior to the cold challenge, a baseline image of both hands (dorsal aspect) was taken with LSCI and both thermal cameras. As required for LSCI imaging, all images were acquired in low-lit rooms. The patient's hands were placed on a black, thermally insulated surface (one metre away from the thermal cameras and 70cm (+/- 5cm) from the LSCI). Small sticky dots were used to mark the location of each finger at baseline. Both hands were

nitrile gloved and immersed to the metacarpophalangeal joints for 1 minute into cooled water; 2 standard containers at 15 ± 1°C (measured by calibrated thermometer) one on either side of the patient. After the cold challenge, gloves were removed and hands returned to their original position on the insulating surface, secured by double sided sticky tape to avoid movement between images. Reperfusion/rewarming after cold challenge was imaged simultaneously by LSCI at 15 frames per minute and thermography (at 4 frames per minute) for 15 minutes (i.e contemporaneous measurement for 15 minutes post cooling). Mobile phone thermography did not allow for continuous video images to be obtained from which data could be extracted and thus single images were taken at set time points; baseline, 0 and 15 minutes after cold challenge. At the end of the 15 minutes one extra image was taken for LSCI and standard thermography to allow the gradient of the last data point to be calculated; thus a total of 225 images/scans were obtained for LSCI, 61 for thermography and 3 for mobile phone thermography during the 15 minutes of measurement. Analysis was performed in Moor Instruments Laser Perfusion Imager software version 4.0 for LSCI, and Research IR max version 4.2, [FLIR, Sweden] for standard and mobile thermography. Patients completed an RCS (0-10) at each visit ('RCS on the day'), measuring the severity and impact of their RP for that day (24).

The cold challenge was repeated one day later (Day 2) as close as possible to the same time of day to minimise variation due to circadian rhythms (27). The repetition over 2 *consecutive* days (i.e. approximately 24 hours) minimised any variations within individuals over time (e.g. menstrual cycle effects) and seasonal variation in weather (28). Five centres had one observer, one centre had 2. Each examiner re-examined same subject on days one and two.

(for example at the central site one observer imaged 60 patients, twice, on consecutive days) Figure 2 shows the study design.

Image analysis: summary measures of response

Image analysis was carried out locally by an internal non-blinded observer at each centre. These were the same observers that had taken the images. Regions of interest (ROIs, Figures 1b and 2b and c), were highlighted in the baseline (pre cold challenge) image and in sequential images for 15 minutes post cooling. The distal dorsal difference (DDD, measurement difference between dorsum and finger, (29,30) (with subscripted L, T or M for LCSI, thermography and mobile phone thermography respectively where applicable) was calculated for each finger at baseline. In the sequential images the ROIs were confined to the 8 distal phalanges. The area under the reperfusion/rewarming curve (AUC) for each finger was calculated manually, not by automation, (Figure 3 [standard thermography],3) from 61 post challenge images as well as the maximum blood flow/temperature after rewarming (MAX), and the gradient in the first 2 minutes (GRAD). Data were averaged for all fingers as in previous studies (16). For mobile phone thermography DDD was taken from the first of 3 images and AUC approximated by averaging over the latter two images. Analysis took less than one hour per participant, per visit.

Saved images and ROI local analysis data were also analysed by the blinded central observer. Mobile phone thermography image analysis was carried out only at the central site.

Feasibility

Feasibility was assessed at the end of the study by clinical scientist/technician opinion (ease of use and analysis 0-10 [0=difficult], preference of LCSI or thermography [standard or mobile] for acquiring and analysing images).

Room temperature measurement

A prerequisite of the cold challenge, and thus centre participation, was a temperature-controlled room at each centre. All measurements were taken in a temperature controlled room (aimed at 23 ± 2°C). Room temperatures were recorded with data monitors (TinyTag, Gemini Data Loggers, UK) to assess the impact of temperature on measurements, with an interest in examining whether reliability could be improved by achieving greater temperature control.

Edge effects from LCSI

It became apparent when the study began that the blood flow appeared to be lower at the edges of the LSCI images than in the centre. This implied that the distribution of the laser light across the hands was not equal, with less light incident towards the edges than at the centre of the image. If true then the consequence of this would be artificially lower value perfusion for little fingers (edge of image, Figure 4) compared to index fingers (centre of image) in the LSCI images. Thus this was investigated further as detailed below.

Statistical analysis

Analysis was performed using R version 3.2.3 (31).

Based upon a previous thermography study (16) 180 patients would allow us to estimate reliability to within 0.05. A full discussion of the sample size calculation and other aspects of the statistical analyses are provided in the supplementary material (extended statistical analysis).

Test-Retest Reliability of the techniques: Intraclass correlation coefficients (ICCs) were obtained using linear mixed effects models with each summary measure included as a dependent variable and with centre as a fixed effect.

Reliability between observers: The data over both visits for each patient were averaged and the resulting averages between central and centre-specific observer compared by calculating the difference and 95% CI for the paired means (Table S2). It is not possible to calculate a valid inter-observer ICC from this data, which would require at least some of the participants to have travelled to all sites for imaging and a large subset of images analysed by all observers (32).

Validity of the techniques: Convergent validity (one aspect of construct validity) was assessed using bivariate linear mixed models including fixed centre terms and separate random patient intercepts for 1) LSCI and standard thermography and 2) standard and mobile thermography. We estimated the latent correlation (which would be equal to one if

the techniques measured the same construct). For clarity, the mathematical representation of this joint model is provided in the supplementary material [statistical analysis (protocol)].

A post-hoc analysis was conducted where the responses to the RCS corresponding to the study day were related to the measurements, using linear mixed models.

Feasibility: Descriptive statistics were used to summarise the feasibility data.

Room temperature: The mean temperature for each patient visit was added to the mixed effects models for each summary measure. ICCs were recalculated, and compared with the previously calculated estimates.

Edge effects: This was investigated in a post-hoc analysis by calculating the trend across fingers for LSCI measurements and comparing these to thermography. Linear mixed models were used to assess any linear trends in the measurements from the index finger to the little finger. Fingers were numbered.Finger-level summary measures of response were then regressed on finger number for both LSCI and thermography; this linear approximation was crude but sufficient. Random intercept and slope terms were included to account for the fact that there was variation from patient to patient in these trends not attributable to the imaging techniques. Measurements were standardised prior to analysis, thereby allowing comparison to be made between LSCI and thermography.

RESULTS

159 patients were recruited (60 from the central centre, 16-20 from each of the others): 157 (99%) fulfilled the 2013 American College Rheumatology/European League Against Rheumatology criteria (34); median age 63.3 IQR (53.8-69.5) years; 123 (77%) lcSSc;

duration since first non-Raynaud's symptom 9.6 (4.5-17.4) years; 146 (93%) were on vasodilators (62 calcium channel blockers, 27 ACE inhibitors, 27 angiotensin II receptor antagonist, 22 Phosphodiesterase-5 inhibitor, 4 endothelin Receptor Antagonist, 1 Nitrates. 35 were on one or more vasodilator); 4 (3%) had previously had finger surgical debridements; 5 (3%) had previously had amputations; 30 (19%) had experienced ulcers in the last year.

Test-Retest Reliability of the techniques: There was at least moderate to substantial reliability for $DDD_{L,M,T}$ $AUC_{L,T,M}$ and $MAX_{L,T}$. $GRAD_{L,T}$ were fair to substantial (Table 1 (35)). A value of 0.7 could be considered high such that both MAX and AUC display strong convergence. Strength of reliability: 0.00 to 0.20 = slight; 0.21 to 0.40 = Fair; 0.41 to 0.60 = Moderate; 0.61 to 0.80 = Substantial; 0.81 to 1.00 = Almost perfect as per (33) although these classifications are to some extent arbitrary and should be treated as a rough guide.

Reliability between observers: The data for each visit, observer and centre are displayed in Supplementary Table S2 and additionally at patient-level using ladder plots (Figure S1). If the measures were perfectly reliable the subplot for each centre would look like two identical ladders (but it is not expected that the plots will be identical between centres). Our data suggest that there were systematic differences between the central observer and centre 2 (and possibly centre 3) in extracting data from LSCI images. For thermography, agreement between the central and local observer was generally high for all centres, albeit with a large discrepancy for several patients for one of their visits.

Validity of the techniques: The latent correlation (95% CI) for LSCI and thermography (i.e. evidence that LCSI and standard thermography measure the same construct, in this case blood returning to the finger) was DDD: 0.65 (0.50 to 0.79); AUC: 0.94 (0.87 to 1.00) and

MAX: 0.87 (0.77 to 0.95); but GRAD only 0.52 (0.33 to 0.70), Table 1. High latent correlation is indicative of convergent validity. A value of 0.7 could be considered high such that both MAX and AUC display strong convergent validity. Correlation between mobile thermography and standard thermography was also very high; 0.98 (0.94 to 1.00) for AUC and 0.90 (0.79 to 0.97) for DDD, Latent correlation between LSCI and FLIR was 0.86 (0.74 to 0.97) for AUC and 0.49 (0.29 to 0.66) for DDD. (Table 1).

With the exception of some weak evidence of decreasing DDD_T with increasing RCS (-0.15 DDD_T for a one point increase in RCS, on average, SE = 0.07) we found no evidence of correlation between the summary measures and RCS.

Feasibility: Standard thermography was deemed to be more feasible than LSCI (see discussion). The proportion of raters giving a score of 7 or above for ease of use (0=difficult–10=easy) was 50%for LSCI; 75%for standard thermography and 38%for mobile phone thermography. Ease of analysis was rated as 7 or above by 25% for LSCI and 50%for standard thermography. The number of centres preferring LSCI to thermography was 1 for acquiring and 1 for analysing images; preferring standard thermography 3 (acquiring) and 4 (analysing). The remaining centres showed no preference.

Room temperature: When included as a covariate, temperature was not associated with any of the summary measures as measured by either LSCI or thermography. Additionally, the ICCs were not affected by the inclusion of temperature in the analysis. This does not mean that a regulated room temperature is not important but that small changes in temperature are acceptable (supplementary Table S3).

Edge effects: Moving from the thumb to the little finger, all of the AUC, MAX and GRAD trends were in the opposite direction for the two modalities, with a decrease for LSCI and an

increase for thermography (supplementary Table S4). Both estimates for the DDD are positive, but this was attenuated for LSCI. This is consistent with an edge effect artificially producing lower values for the little fingers with LSCI. The cause of the edge effect was identified as the distribution of the light over the imaging area, due to LSCI being used at the upper limit of the suggested imaging distance in order to fit both hands into the imaging area. The data indicates that care must be taken to understand the variations over the field of view so that these can be accounted for; decreasing the field of view would minimise these results in future studies.

DISCUSSION

To date laser speckle imaging techniques and thermography have been insufficiently studied as outcome measures in clinical trials. Those studies in which they have been included show very little consistency in terms of protocol design (24, 36-38), choice of dynamic challenge and extracted outcome measures, making it difficult to compare results between studies or establish a standard protocol. The main finding of our study is that reliability of both LSCI and thermography were sufficiently high (AUC and MAX) for use as study outcome measures. The reliability of MAX_T was slightly superior to MAX_L. Other than this, there were no substantive differences in reliability between the two techniques.

AUC_M and DDD_M showed adequate reliability for use as outcome measures. Moreover, there was strong correlation between mobile phone and standard thermography data. The mobile phone thermography was added at a late stage in this project (since it had only just come on the market). Our reason for including it was primarily for feasibility assessment. While it is

clear that further work is required to validate mobile thermography, the performance in the present study is highly encouraging because as a low-cost tool it could potentially be readily available for widespread use amongst rheumatologists.

Although not our primary objective, we examined differences between observers. Systematic differences between observers at different centres would not be particularly problematic for a multicentre randomised controlled trial (RCT), provided randomisation was stratified by centre. We note that this should be the default for any multicentre trial, since differences between centres may otherwise bias the estimated treatment effect. This is particularly true in small populations, since simple randomisation is less likely to produce balance within centres. Standardised training would reduce measurement variation across centres and centralised blinded extraction and analysis of LSCI data might also minimise variation by removing multi-observer differences in an RCT setting. Given the small sample size at each centre we are unable to determine whether truly systematic differences were observed. Ideally, a study to assess inter-observer reliability would involve participants having images analysed by all observers.

Convergence between the techniques was shown to be very high for AUC and MAX (particularly for AUC). This provides evidence that the same underlying construct is being measured when using these summaries of response. Convergence appeared to be weaker (although still moderate) for DDD. Convergence was weakest for GRAD which may reflect a lag between tissue re-perfusion (measured with LSCI) being translated into skin re-warming (measured using thermography) during the 2 minutes immediately following cold challenge.

Since there is no gold standard to compare either imaging technique to, and we are comparing two techniques that measure perfusion by very different methods (skin temperature and a measure of red blood cell concentration and speed by light) it is possible to measure convergence between these techniques for validity (25). It would be unlikely for these two techniques to converge whilst also being poor measures since they would both have to be deficient in distinct but very specific ways so as to bring the erroneous observations into alignment. Therefore we can conclude in this instance that their convergence implies validity.

The OMERACT review of 2003 (18) assessed the validity of several non-invasive techniques as possible objective outcome measures but none was deemed ready for use in clinical trials. These included nailfold capillaroscopy, a well-established diagnostic technique is now included in the diagnostic criteria (34) to differentiate primary and secondary Raynaud's phenomenon. The microscopy technique allows visualisation of cutaneous capillaries at the nailbed and identification of structural change characteristic of SS. This is not a substitute though for functional measures of flow (although functional flow and oxygenation have recently been reported). Plethysmography allows the change in vascular volume to be measured (i.e. detection of a pulse) in combination with cold challenge. The technique can measure full fields in the same ways as laser speckle but remains unvalidated. There was no relationship between the summary measures and the RCS on the day of the study visits, for either LSCI or thermography. Patient-centred outcome measures are crucial for evaluating the effectiveness (rather than just the efficacy) of treatments. However, patient-centred outcomes often comprise more 'noise' compared to more objective measures of response, and therefore necessitate larger sample sizes to ensure adequate power in clinical trials. For

small populations, there is therefore a tension between direct relevance to patients and feasibility of conducting a trial. One solution may be to power studies on the basis of objective measures such as those considered here, and to additionally (and consistently) report patient-centred outcomes to facilitate an eventual meta-analysis. Another might be to seek confirmatory evidence for the vasodilatory potential of candidate interventions using objective measures before proceeding to larger phase 3 clinical trials

The relationship between two measures is limited by the reliability of each (39). While relative stability of RCS has been observed between baseline and follow-up in clinical trials/studies (38,40), there has been little work formally assessing its intra-individual reliability.

Regarding feasibility, comments were made regarding LSCI and its sensitivity to movement, vibrations and lighting, indicating the importance of environmental conditions. For the mobile phone thermography, present limitations include battery life (LSCI and thermography were mains/long-life battery powered), fixed focussing distance and lack of analysis for video images, as well as mounting difficulties; however, if the mobile and standard thermography correlation can be replicated in future studies, these limitations may be acceptable in light of the lower cost and ambulatory (convenient) nature of the technique. When comparing feasibility of LSCI versus thermography, it should be noted that most centres were familiar with thermography but not LCSI, and that this may have influenced assessment of feasibility.

One limitation of the study was that we did not recruit the planned number of participants, due to a seventh centre not participating as planned. However, the study was designed to be robust to under-recruitment. Although the 95% CIs for our estimates are wider than they

would have been had the target been met, we were still able to demonstrate good enough reliability and convergent validity of AUC and MAX, to observe differences indicative that that the performance of DDD was weaker, and to show that the performance of GRAD was relatively poor.

In conclusion, our design was relatively pragmatic, with the aim of establishing the performance of the different techniques as they would be employed in a multicentre clinical trial. Our study successfully established a working group of tertiary-SSc centres, and together the group developed a consensus calibration and cold challenge protocol. The summary measures AUC and MAX both displayed good reliability and strong convergent validity. There was a possible advantage of thermography in relation to the reliability of MAX, although this was not definitive. We found evidence of edge effects when using LSCI although our summary measures appeared to be quite robust to these in relation to reliability, perhaps suggesting that these effects were fairly consistent (methods and results discussed in supplementary data). The study has also confirmed that small variations in room temperature are acceptable and that, subject to further validation, mobile phone cameras may be a suitable, affordable, highly portable alternative to more expensive standard imaging equipment (although mobile phones are battery operated and with less functionality [at present] than larger thermal cameras). The mobile phone data obtained in this study will facilitate the design of future validation studies into mobile phone thermography-derived outcome measures. Although the design precluded formal assessment of inter-observer reliability there was a suggestion of systematic differences between the central observer and observers at some centres, highlighting the importance of

image analysis training and potentially a role for centralised or automated image analysis. For multicentre RCTs, we would also recommend that where possible/appropriate, randomisation be stratified by centre to balance any centre effects and prevent bias.

In summary LSCI and thermography should now be incorporated as secondary outcomes in upcoming treatment efficacy trials. This will allow an assessment of responsiveness to treatment as well as longitudinal validity. The present study leads us to recommend the summary measures AUC and MAX, measured using both thermography and LSCI (but especially using thermography), as suitable outcome measures for RCTs in SSc-related RP.

ACKNOWLEDGEMENTS

This study was funded by an Arthritis Research UK Clinical Studies Grant [20656] and an unrestricted educational grant from Actelion Pharmaceuticals Ltd. This study was supported by Arthritis Research UK grant number (Centre for Epidemiology) 20380 and the Manchester Academic Health Sciences Centre (MAHSC). The authors would like to acknowledge the assistance given by IT Services and the use of the Computational Shared Facility at The University of Manchester.

We are grateful to the trial steering and data monitoring committee members: Dr Mohammad Akil, Prof David D'Cruz and the late Prof Peter Wells.

We also thank Dipa Ghedia at the London site and Sook Eng at the Leeds site for patient recruitment, Darren Hart at the Bath site for imaging and analysis and Anita Furlong and Tracey Drayton at the Cambridge site for imaging.

We wish to thank the UK Scleroderma Study Group for their advice and support in the development and running of this study and Moor Instruments and Thermal Vision Research for their advice and training.

REFERENCES

1. Walker UA, Tyndall A, Czirjak L, Denton C, Farge-Bancel D, Kowal-Bielecka O, et al. Clinical risk assessment of organ manifestations in systemic sclerosis: a report from the EULAR Scleroderma Trials and Research group database. Ann Rheum Dis 2007;66:754-63.
2. Willems LM, Kwakkenbos L, Leite CC, Thombs BD, van den Hoogen FH, Maia AC et al. Frequency and impact of disease symptoms experienced by patients with systemic sclerosis from five European countries. Clin Exp Rheumatol 2014;32:S-88-93
3. Rodnan GP, Myerowitz RL, Justh GO. Morphological changes in the digital arteries of patients with progressive systemic sclerosis (scleroderma) and Raynaud's phenomenon. Medicine 1980;59:393–408.
4. Herrick A. Diagnosis and management of scleroderma peripheral vascular disease. Rheum Dis Clin North Am 2008;34:89-114.
5. Steen VD, Powell DL, Medsger TA. Clinical correlations and prognosis based on serum autoantibodies in patients with systemic sclerosis. Arthritis Rheum 1988;31:196-203.
6. Della Rossa A, Valentini G, Bombardieri S, Bencivelli W, Silman AJ, D'Angelo S, et al. European multicentre study to define disease activity criteria for systemic sclerosis. Clinical and epidemiological features of 290 patients from 19 centres. Ann Rheum Dis 2001;60:585-91.
7. Ferri C, Valentini G, Cozzi F, Sebastiani M, Michelassi C, La Montagna G, et al. Systemic sclerosis: demographic, clinical and serologic features and survival in 1,012 Italian patients. Rheumatol 2002;81:139-53.
8. Mawdsley AH. Patient perception of UK scleroderma services – results of an anonymous questionnaire. Rheumatol 2006;45:1573.

9. Tiev KP, Diot E, Clerson P, Dupuis-Siméon F, Hachulla E, Hatron PY, et al. Clinical features of scleroderma patients with or without prior or current ischemic digital ulcers: post-hoc analysis of a nationwide multicenter cohort (ItinerAIR-Sclerodermis) J Rheum 2009;36:1470-6.
10. Khimdas S, Harding S, Bonner A, Zummer B, Baron M, Pope J; et al. Associations with digital ulcers in a large cohort of systemic sclerosis: results from the Canadian Scleroderma Research Group Registry. Arth Care Res 2011:63;142-9.
11. Ennis H, Vail A, Wragg E, Taylor A, Moore T, Murray A, et al. A prospective study of systemic sclerosis-related digital ulcers: prevalence, location, and functional impact. Scand J Rheumatol. 2013;42:483-6.
12. Ennis H, Hughes M, Anderson ME, Wilkinson J, Herrick AL. Calcium channel blockers for primary Raynaud's phenomenon. Cochrane Database of Systematic Reviews 2016. DOI: 10.1002/14651858.CD002069.pub5.
13. Herrick AL. Raynaud's phenomenon (secondary). BMJ Clinical Evidence (Online) 2008; 1125.
14. Stewart M, Morling JR. Oral vasodilators for primary Raynaud's phenomenon. Cochrane Database of Systematic Reviews 2012, Issue 7. Art. No.: CD006687.
15. Garcia de la Pena Lefebvre P, Nishishinya MB, Pereda CA, Loza E, Sifuentes Giraldo WA, Román Ivorra JA, et al. Efficacy of Raynaud's phenomenon and digital ulcer pharmacological treatment in systemic sclerosis patients: a systematic literature review. Rheumatol Int 2015;35:1447-1459.
16. Murray AK, Moore TL, Manning JB, Taylor C, Griffiths CE, Herrick AL. Noninvasive imaging techniques in the assessment of scleroderma spectrum disorders. Arthritis Rheum 2009;61:1103-11.

17. Pauling JD, Shipley JA, Raper S, Watson ML, Ward SG, Harris ND, et al. Comparison of infrared thermography and laser speckle contrast imaging for the dynamic assessment of digital microvascular function. Microvasc Res 2012;83:162-7.

18. Merkel PA, Clements PJ, Reveille JD, Suarez-Almazor ME, Valentini G, Furst DE, et al. Current status of outcome measure development for clinical trials in systemic sclerosis. Report from OMERACT 6. J Rheumatol 2003;30:1630-47.

19. Herrick AL. Secondary Raynaud's phenomenon. BMJ Clinical Evidence 2008;09:1125.

20. Cerinic M, Denton CP, Furst DE, Mayes MD, Hsu VM, Carpentier P, et al. Bosentan treatment of digital ulcers related to systemic sclerosis: results from the RAPIDS-2 randomised, double-blind, placebo-controlled trial. Ann Rheum Dis 2011;70:32-8.

21. Fava A, Wung PK, Wigley FM, Hummers LK, Daya NR, Ghazarian SR, et al. Efficacy of Rho kinase inhibitor fasudil in secondary Raynaud's phenomenon. Arthritis Care Res 2012;64:925-9.

22. Khanna D, Denton CP, Merkel PA, Krieg T, Le Brun FO, Marr A, et al. Effect of Macitentan on the Development of New Ischemic Digital Ulcers in Patients With Systemic Sclerosis: DUAL-1 and DUAL-2 Randomized Clinical Trials. JAMA. 2016;315:1975-88.

23. Seibold JR, Wigley FM, Schiopu E, Denton CP, Silver RM, Steen VD, et al. Digital ulcers in SSc treated with oral treprostinil: a randomized, double-blind, placebo-controlled study with open-label follow-up. J Scleroderma Relat Disord 2017; 2: 42-9.

24. Merkel PA, Herlyn K, Martin RW, Anderson JJ, Mayes MD, Bell P, et al. Measuring disease activity and functional status in patients with scleroderma and Raynaud's phenomenon. Arthritis Rheum 2002;46:2410-20.

25. Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use. Oxford University Press, USA; 2014 Oct 30.
26. Clark S, Dunn G, Moore T, Jayson M 4th, King TA, Herrick AL. Comparison of thermography and laser Doppler imaging in the assessment of Raynaud's phenomenon. Microvasc Res 2003;66:73–6.
27. Houben AJ, Slaaf DW, Huvers FC, de Leeuw PW, Nieuwenhuijzen Kruseman AC, Schaper NC. Diurnal variations in total forearm and skin microcirculatory blood flow in man. Scand J Clin Lab Invest 1994;54:161-8.
28. Bartelink ML, Wollersheim H, Theeuwes A, van Duren D, Thien T. Changes in skin blood flow during the menstrual cycle: the influence of the menstrual cycle on the peripheral circulation in healthy female volunteers. Clin Sci (Lond) 1990;78:527-32.
29. Clark S, Hollis S, Campbell F, Moore T, Jayson M, Herrick A. The "distal-dorsal difference" as a possible predictor of secondary Raynaud's phenomenon. J Rheumatol 1999;26:1125-8.
30. Anderson M, Moore T, Lunt M, Herrick AL. The 'distal-dorsal difference': a thermographic parameter by which to differentiate between primary and secondary Raynaud's phenomenon. Rheumatol 2007;46:533-8.
31. R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
32. Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Physical therapy 1994 Aug 1;74(8):777-88.

33. LeRoy EC, Black C, Fleischmajer R, Jablonska S, Krieg T, Medsger TA Jr, et al. Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. J Rheumatol 1988;15:202–5.

34. van den Hoogen F, Khanna D, Fransen J, Johnson SR, Baron M, Tyndall A, et al. 2013 classification criteria for systemic sclerosis: an American College of Rheumatology/European League against Rheumatism collaborative initiative. Arthritis Rheum 2013 Nov;65(11):2737-47.

35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-174.

36. Pauling J, Shipley J, Harris N, McHugh NJ. Use of infrared thermography as an endpoint in therapeutic trials of Raynaud's phenomenon and systemic sclerosis. Clin Exp Rheumatol 2012;30:S103-15.

37. Allen J, Howell K. Microvascular imaging: techniques and opportunities for clinical physiological measurements. Physiol Meas 2014;35:R91-R141.

38. Pauling JD, Shipley JA, Hart DJ, McGrogan A, McHugh NJ. Use of Laser Speckle Contrast Imaging to Assess Digital Microvascular Function in Primary Raynaud Phenomenon and Systemic Sclerosis: A Comparison Using the Raynaud Condition Score Diary. J Rheumatol 2015;42:1163-8.

39. Fleiss JL. Reliability of Measurement. Ch.1 in Design and Analysis of Clinical Experiments. Volume 73 of Wiley Classics Library. John Wiley & Sons, 2011.

40. Gladue H, Maranian P, Paulus HE, Khanna D. Evaluation of test characteristics for outcome measures used in Raynaud's phenomenon clinical trials. Arthritis Care Res 2013;65:630-6.

LEGENDS

Figure 1: a) Photograph of the imaging equipment set-up to allow simultaneous imaging, showing LSCI and standard and mobile thermography; b) Baseline image of hands taken with standard thermography showing distal dorsal difference (DDD) regions of interest (ROI), fingers cooler than dorsum; c) An example of hands imaged by mobile phone thermography at 0 mins post cold challenge, fingers cooler than dorsum (scale unavailable for image due to software); d) An example of hands (same subject as in b)) imaged by the standard thermal camera undergoing rewarming (from left to right, 0 mins after cooling, 7.5 mins and 15 mins), scale to the right 20-37 ^{o}C refers to b and d.

Figure 2: Study design demonstrating images taken contribute to convergent validity, test-retest and intra-observer differences.

Figure 3: Example of a rewarming curve (for one hand measured with standard thermography), temperature vs rewarming time, one line for each finger (ROI as per b)). Annotation: area under the curve (AUC) bounded by green rewarming line (for middle finger) and 2 green dotted lines, maximum temperature (MAX), shown with red dotted line and arrow and gradient in the first 2 mins (GRAD) shown in enlarged box, represented by red dotted line for index finger (red rewarming curve).

Figure 4: a) LSCI reperfusion graphs (perfusion [flux (proportional to the product of the average speed of the blood cells and their number concentration, expressed in arbitrary 'perfusion units'), vs. time]) for 8 digits (ROI 1-4 and 6-9 in 2b) and 2 dorsa (ROI 5 and 10; b) Example flux (ie perfusion map) image showing ROIs marked (see Figure 1); c) photographic image of hands showing ROIs.

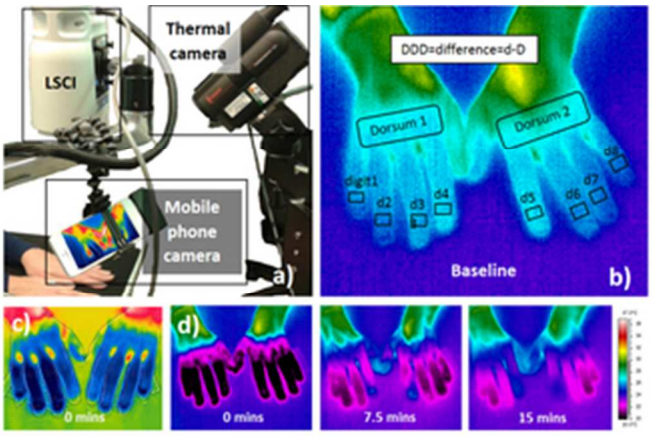


Figure 1: a) Photograph of the imaging equipment set-up to allow simultaneous imaging, showing LSCI and standard and mobile thermography; b) Baseline image of hands taken with standard thermography showing distal dorsal difference (DDD) regions of interest (ROI), fingers cooler than dorsum; c) An example of hands imaged by mobile phone thermography at 0 mins post cold challenge, fingers cooler than dorsum (scale unavailable for image due to software); d) An example of hands (same subject as in b)) imaged by the standard thermal camera undergoing rewarming (from left to right, 0 mins after cooling, 7.5 mins and 15 mins), scale to the right 20-37 oC refers to b and d.

28x18mm (300 x 300 DPI)

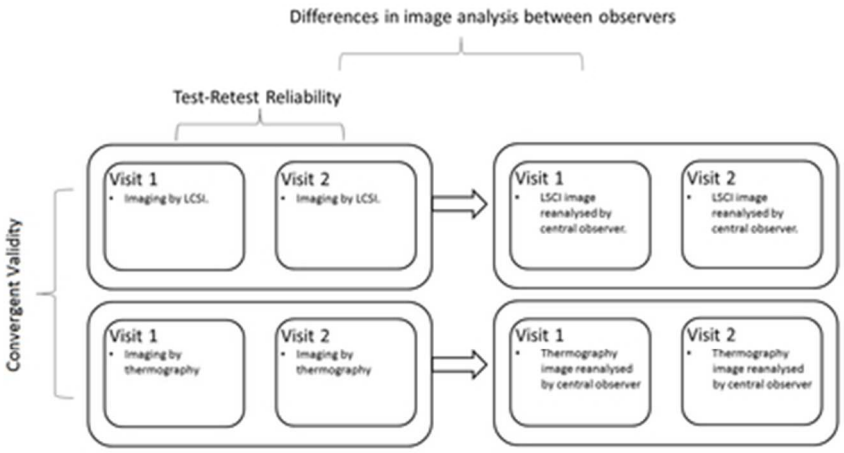


Figure 2: Study design demonstrating images taken contribute to convergent validity, test-retest and intra-observer differences.

36x19mm (300 x 300 DPI)

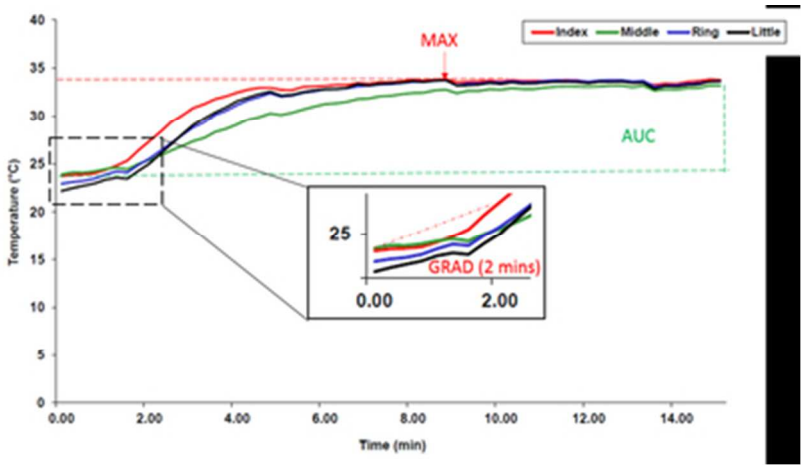


Figure 3: Example of a rewarming curve (for one hand measured with standard thermography), temperature vs rewarming time, one line for each finger (ROI as per b)). Annotation: area under the curve (AUC) bounded by green rewarming line (for middle finger) and 2 green dotted lines, maximum temperature (MAX), shown with red dotted line and arrow and gradient in the first 2 mins (GRAD) shown in enlarged box, represented by red dotted line for index finger (red rewarming curve).

34x19mm (300 x 300 DPI)

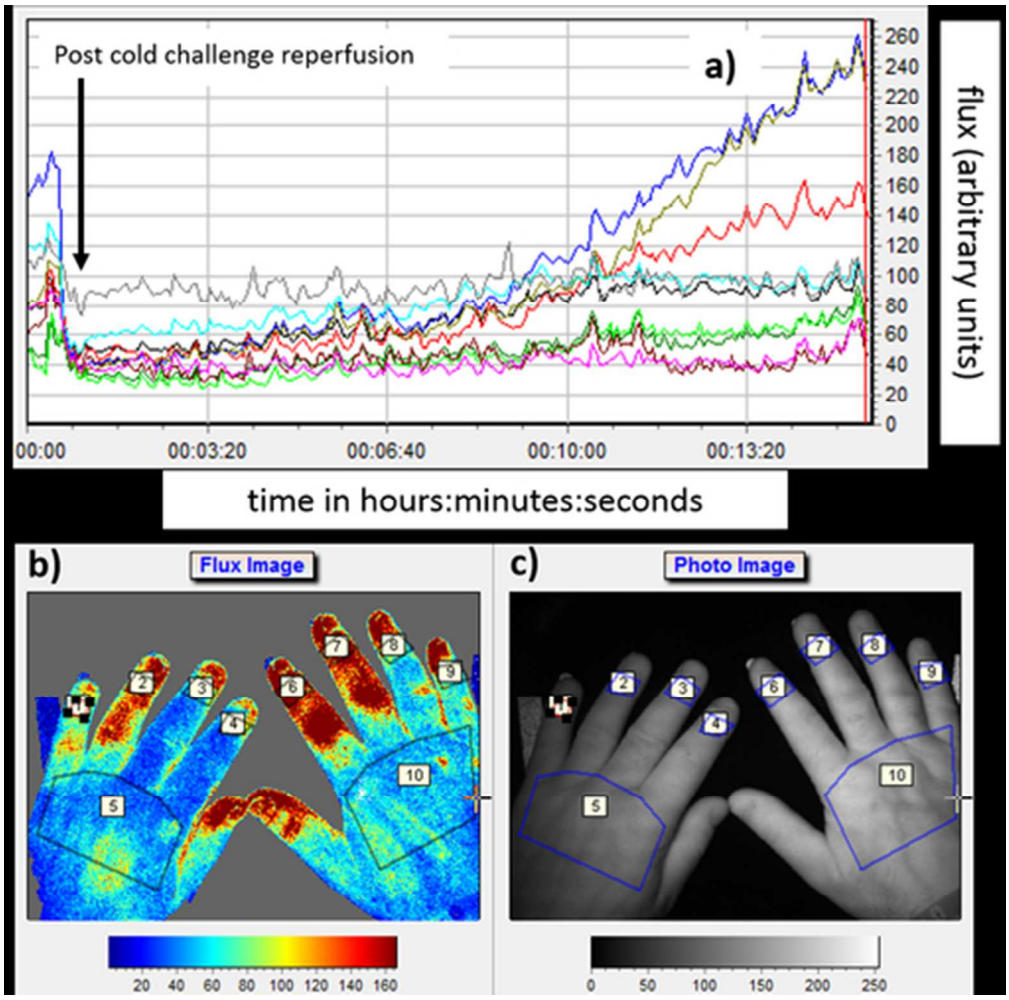


Figure 4: a) LSCI reperfusion graphs (perfusion [flux (proportional to the product of the average speed of the blood cells and their number concentration, expressed in arbitrary 'perfusion units'), vs. time]) for 8 digits (ROI 1-4 and 6-9 in 2b) and 2 dorsa (ROI 5 and 10; b) Example flux (ie perfusion map) image showing ROIs marked (see Figure 1); c) photographic image of hands showing ROIs.

43x42mm (300 x 300 DPI)

Table 1: Reliability and validity of Laser speckle contrast imaging (LCSI) and thermography (standard and mobile phone) in patients with SSc-related RP.

	Test-retest reliability [ICCs and 95% confidence intervals (CIs)]			Comparison of reliability between LSCI and standard thermography [difference in reliability]	Validity [estimated latent correlation between LSCI and standard thermography and between standard and mobile phone thermography]	
Summary measure	LSCI (N=159)	Standard thermography (N=159)	Mobile phone thermography (N=141**)	Difference† in ICCs: LSCI minus Thermography	Correlation of LCSI and standard thermography	Correlation of standard and mobile phone thermography
Distal dorsal difference DDD	0.67 (0.56 to 0.77)	0.58 (0.43 to 0.71)	0.61 (0.51 to 0.73)	0.08 (-0.05 to 0.25)	0.65 (0.50 to 0.79)	0.90 (0.79 to 0.97)
Area under reperfusion /rewarming curve Log(AUC)	0.67 (0.54 to 0.76)	0.68 (0.58 to 0.80)	0.61 (0.51 to 0.72)*	-0.01 (-0.17 to 0.11)	0.94 (0.87 to 1.00)	0.98 (0.94 to 1.00)
maximum perfusion /temperature Log (MAX)	0.64 (0.52 to 0.75)	0.72 (0.64 to 0.81)	NA	-0.09 (-0.21 to 0.03)	0.87 (0.77 to 0.95)	NA
gradient over first 2	0.46 (0.40	0.56 (0.40 to	NA	-0.09 (-0.24 to	0.52 (0.33 to	NA

minutes GRAD	to	0.74) 0.69)		0.18)	0.70)

N=number of participants. Data have been averaged over 8 digits.

*AUC for mobile phone approximated by mean of 2 frames, post cold challenge);

**141 data sets were available for the mobile phone thermography (n=18 missing due to technical fault at two centres).

† Differences between the point estimates for LSCI and standard thermography (test-reliability columns, 3 and 4).

Supplementary Table S1: Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
Signed informed consent.	Primary RP or secondary RP due to aetiology other than SSc.
RP defined as a history of digital cold sensitivity associated with colour changes (cyanosis and pallor). SSc as diagnosed by an experienced rheumatologist and fulfilling either the American College of Rheumatology criteria for SSc (32) or the criteria for early disease as defined by LeRoy and Medsger et al (40).	Inability to undergo LSCI or thermography due to active digital ulcers or inability to extend fingers sufficiently or any other skin features that would lead to.
Male and female patients aged ≥ 18 years at screening.	Any significant organ involvement or concomitant condition which, in the opinion of the investigator, would make in unwise for the patient to participate in the study.
Stable vasoactive medication: doses stable for at least 1 month prior to Visit 1 and until Visit 2.	Any disorder limiting the ability to provide informed consent or to comply with study requirements.

	Diabetes mellitus which affects the microcirculation.
	Females who were breast feeding or pregnant.
	Smokers; the use of nicotine patches was not allowed as these cause vasoconstriction.
	Change of vasoactive medication within 1 month prior to Visit 1 or planned change of medication between the study visit dates.
	Treatment with prostacyclin (epoprostenol) or prostacyclin analogs (i.e., iloprost, treprostinil) within 1 month prior to Visit 1.
	Local treatment (digits) with botulinum toxin type A within 1 month prior to Visit 1.
	Topical administration of nitrates within 1 week prior to Visit 1.
	Treatment with vasoconstrictive drugs (e.g., ergot derivatives, triptans) within 1 week prior to Visit 1.
	Surgical sympathectomy of digit within 3 months prior to Visit 1.

Supplementary Table S2: Summary measures of responses to cold challenge comparison of data taken by central and centre-specific observers for LCSI and standard thermography.

Central observer vs centre specific observer								
Centre	DDD		Log(AUC)		Log(MAX)		GRAD	
	LSCI (arb PU)	Standard therm. (^{o}C)	LSCI (arb PU*time)	Standard therm. (^{o}C*time)	LSCI (arb PU)	Standard therm. (^{o}C)	LSCI (arb PU/time)	Standard therm. (^{o}C/time)
Centre 2	9.0 (-30.0 to 47.8)	0.02 (-0.92 to 0.96)	0.09 (-0.35 to 0.53)	0.02 (-0.07 to 0.10)	0.10 (-0.19 to 0.39)	0.00(-0.09 to 0.08)	-3.36 (-16.2 to 9.43)	-0.08(-0.37 to 0.22)
Centre 3	48.6 (17.3 to 79.9)	0.18 (-0.46 to 0.82)	0.25 (-0.08 to 0.58)	0.00 (-0.06 to 0.07)	0.22 (-0.04 to 0.49)	0.01(-0.05to 0.08)	1.29(-5.39 to 7.98)	-0.07(-0.46 to 0.31)
Centre 4	-15.7 (48.0 to 63.7)	0.29(-0.76 to 1.34)	-0.15(-0.56 to 0.25)	0.00(-0.07to 0.06)	-0.12(-0.47 to 0.23)	0.00(-0.07 to 0.07)	-0.62(-6.93 to 5.69)	0.05(-0.12 to 0.23)
Centre 5	-12.7(-51.6 to 26.2)	0.11(-0.84 to 1.05)	-0.03(-0.33 to 0.28)	0.00(-0.06 to 0.07)	-0.01(-0.25 to 0.23)	0.00(-0.07to 0.08)	-2.56(-17.9 to 12.8)	-0.06(-0.36 to 0.24)

Centre 6	-5.5(-21.6 to 10.5)	0.06(-0.60 to 0.73)	-0.03(-0.23 to 0.18)	0.01(-0.06 to 0.07)	-0.02(-0.18 to 0.15)	0.01(-0.06 to 0.08)	-0.32(-5.07 to 4.43)	-0.11(-0.48 to 0.26)

For comparison of data between observers: Differences in paired means (95% CI), calculated as central minus centre-specific. Measurements taken by a central blinded observer were compared to the corresponding measurements taken at each centre. The data over both visits for each patient were averaged and the resulting averages between the central observer and the centre-specific observer compared by plotting the data and calculating the difference and 95% CI for the paired means. Centre 1 is excluded from the analysis as there were multiple observers at this site. The exploratory nature of these supplementary analyses should be emphasized.

Supplementary Table S3: Room temperature data for each centre

Centre	One	Two	Three	Four	Five	Six
Room temperature (^{o}C)	21.9	23.6	25.3	23.9	23.2	23.3
	21.6 to 22.3	23.5 to 23.7	24.4 to 24.1	23.6 to 24.1	23.0 to 23.5	22.3 to 24.0
	20.4 to 22.9	23.0 to 24.0	22.6 to 25.6	22.6 to 25.6	21.3 to 24.5	20.2 to 26.9

Summary of temperatures per patient visit (data taken over 45 minutes) at each of six centres. Median, interquartile range, range.

.

Supplementary Table S4: Post hoc analysis of the edge effect observed for LSCI

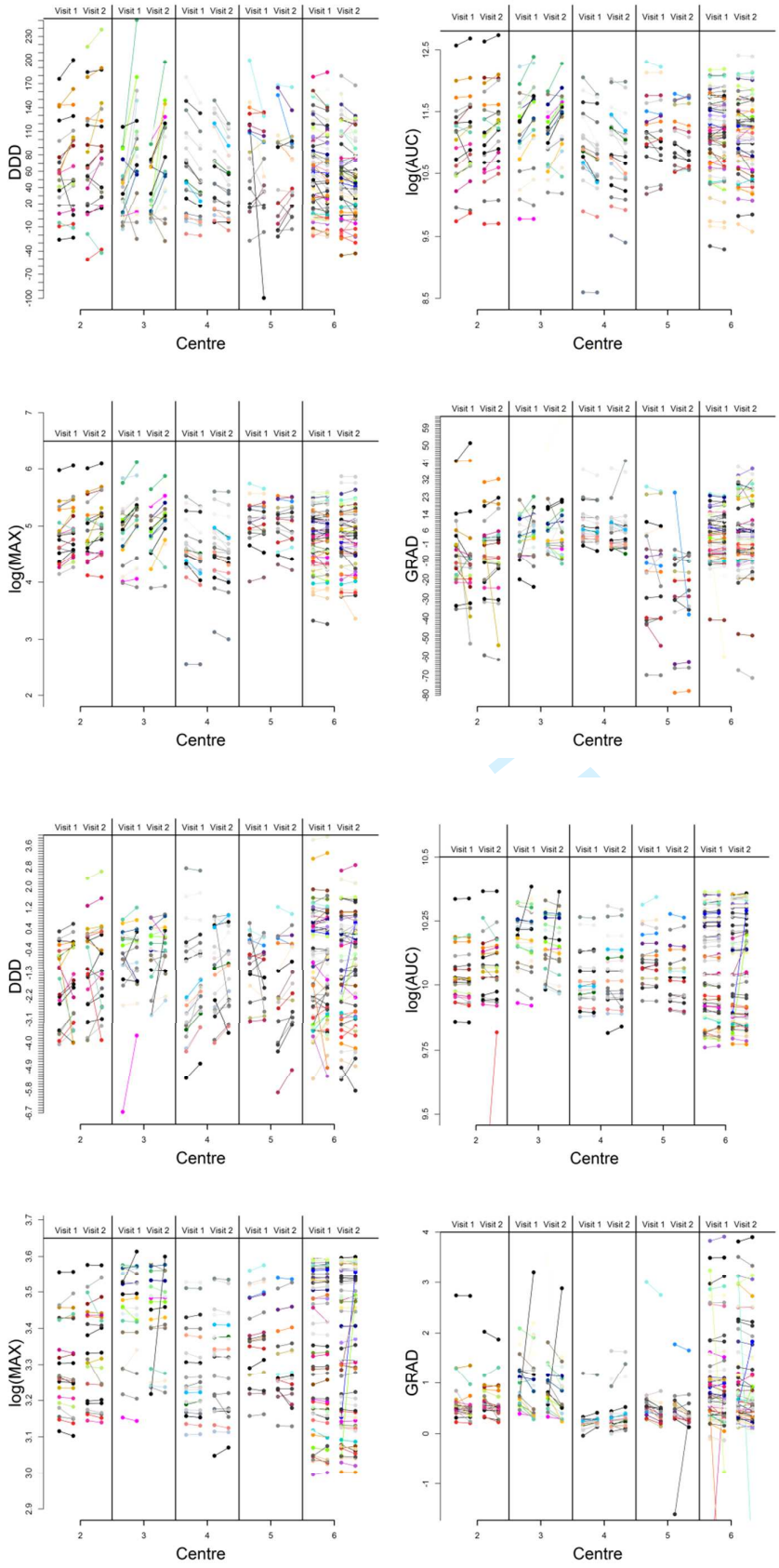
Summary Measure	LSCI	Thermography
Log(AUC)	-0.17(0.02)	0.07 (0.01)
DDD	-0.21 (0.02)	-0.01 (0.02)
Max	-0.19 (0.02)	0.07 (0.01)
Gradient	-0.07 (0.02)	0.04 (0.01)

Estimated trends from index to little fingers, standard error (SEs) from linear mixed models. Units are standard deviations. Positive (negative) trends correspond to increases (decreases) moving across the fingers, with larger values indicating a stronger trend.

Rows show summary measures: Mean (SD); Distal dorsal difference (DDD), area under the reperfusion or rewarming curve (AUC), maximum perfusion or temperature (MAX) and the gradient over the first 2 minutes (GRAD) have been averaged over 8 digits.

a)

b)



Supplementary Figure 1 (S1): Ladder plots showing patient-level summary measures of response measured at each measurement site (first point in each linked pair of points) and again by a universal rater (second point in each linked pair of points). Measurements are shown for each study visit and for a) LSCI and b) thermography.

Supplementary data:

Expansion of statistical analysis of the data

Justification of sample size: The sample size was selected both to estimate reliability precisely and to demonstrate substantial levels of reliability for thermography in the event that this reflected the actual performance of the technique. A reliability estimate for LSCI was not available, although Murray et al. [22] estimated the reliability of laser Doppler to be 0.92. If LSCI were similar, we would obtain a lower 95% confidence limit of 0.89, again demonstrating strong reliability. In the event that the ICCs of the techniques were considerably lower than expected, a sample size of 180 would still allow estimation of the ICCs to a good degree of precision. For instance, if the ICC was as low as 0.65, then 180 patients would allow a 95% confidence interval of overall width 0.17 to be calculated. The interval would then be 0.56 to 0.73. The inference of the study would then be that reliability of the measurement was 'moderate' to 'substantial'. The target sample size was quite robust to substantial under recruitment. For example, we calculated that, on the basis of the figures presented above, a sample size of 120 would result in 95% CIs for thermography and LSCI of overall width 0.13 and 0.06 respectively, which would still represent a good level of precision.

Reliability of the techniques: A fixed effect corresponding to each centre in the study and a patient-specific random intercept were included in the linear mixed effects models. In order to calculate 95% confidence intervals (CIs) for the ICCs a nonparametric bootstrap procedure was performed. In order to compare the reliability of summary measures obtained using LSCI compared to thermography, the difference in ICCs (with bootstrapped 95% CIs) were calculated. AUC and MAX were log transformed due to skewed data prior to

modelling. Data from Mobile phone thermography were collected differently and therefore did not lend themselves to calculating all of the summary measures calculated using LSCI and standard thermography. The mean value and the mean DDD were calculated across eight fingers for each patient for each visit (the former a summary measure approximating AUC), and their reliability across visits was computed.

Reliability between observers: Measurements taken by a central blinded observer were compared to the corresponding measurements taken at each centre. It is not be possible to calculate a valid inter-observer ICC from this data, which would require at least some of the participants to be measured by all of the observers. Without this, the patient-by-observer interaction is not identified [30]. One centre was excluded from this analysis as several observers analysed the images. The exploratory nature of these supplementary analyses should be emphasized.

*Mathematical representation of the model: T*aking one summary measure at a time, for patient i we have one measurement and one replicate obtained using LSCI (x_{1i} and x_{2i}) and one using thermography (y_{1i} and y_{2i}).

For $j = 1,2$ and $i = 1, \dots, n$ we model

$$x_{ji} = \alpha_0 + \alpha_1 centre_i + u_i + \delta_{ji}$$

$$y_{ji} = \beta_0 + \beta_1 centre_i + v_i + \epsilon_{ji}$$

$$\delta_{ji} \sim N(0, \sigma_\delta^2)\ ,\ \epsilon_{ji} \sim N(0, \sigma_\epsilon^2) \text{ and } \begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim MNV\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \omega_u^2 & \rho\omega_u\omega_v \\ \rho\omega_u\omega_v & \omega_v^2 \end{bmatrix}\right)$$

The parameter ρ represents the latent correlation between the two techniques, having separated out the variability due to measurement error. This would be 1 if the techniques were measuring the same underlying construct, and provides evidence of the construct validity of the techniques.

Running head: Multicentre validity and reliability of hand cold challenge as measured by LSCI and thermography

Title: A multicentre study of validity and reliability of responses to hand cold challenge as measured by laser speckle contrast imaging and thermography: outcome measures for systemic sclerosis-related Raynaud's phenomenon

Jack D Wilkinson[1], BSc, MSc

Sarah A Leggett[2], BSc

Elizabeth J Marjanovic[2], BSc, PhD

Tonia L Moore[2], BSc

John Allen[3,4], BSc, PhD

Marina E Anderson[5], FRCP PhD

Jason Britton[6], BSc MSc

Maya H Buch7[7,8], MBChB, FRCP, PhD

Francesco Del Galdo[7,8],MD, PhD

Christopher P Denton[9], PhD, FRCP

Graham Dinsdale[2], MPhys, PhD

Bridgett Griffiths[10], MB ChB, MD, FRCP

Frances Hall[11], MAoxon, FRCP, DPhil

Kevin Howell[9], BSc, MSc, PhD

Audrey MacDonald[3,4], BSc MSc

Neil J McHugh[12], MBChB, MD

Joanne B Manning[13], HNC

John D Pauling[12,14] , BMedSci, PhD, FRCP

Christopher Roberts[1], BSc,PhD

Jacqueline A Shipley[12], BSc, MSc, PhD

Ariane L Herrick[2,15], MD, FRCP

Andrea K Murray[2,16], MPhys, PhD

Corresponding author:

Andrea Murray

Rm C214, Clinical Sciences Building,

Centre for Musculoskeletal Research,

School of Biological Sciences,

Faculty of Biology, Medicine and Health,

Division of Musculoskeletal and Dermatological Sciences,

University of Manchester,

Manchester Academic Health Science Centre,

Salford Royal NHS Foundation,

Manchester,

M13 9PT,

UK.

Andrea.murray@manchester.ac.uk

Tel +44 (0)161 206 1538/1133

Addresses:

1. Centre for Biostatistics,

University of Manchester,

Manchester,

UK

2. Centre for Musculoskeletal Research,

School of Biological Sciences,

Faculty of Biology, Medicine and Health,

Division of Musculoskeletal and Dermatological Sciences,

University of Manchester

Manchester Academic Health Science Centre,

Salford Royal NHS Foundation,

Manchester,

UK

3. Microvascular Diagnostics,

Northern Medical Physics and Clinical Engineering,

Freeman Hospital,

Newcastle upon Tyne,

UK

4. Institute of Cellular Medicine,

Medical School,

Newcastle University,

Newcastle upon Tyne,

UK

5. Institute of Clinical Sciences,

University of Liverpool,

Liverpool,

UK

6. Medical Physics Department,

Leeds Teaching Hospitals NHS Trust,

Leeds,

UK

7. Leeds Institute of Rheumatic and Musculoskeletal Medicine,

Chapel Allerton Hospital,

Leeds,

UK

8. The National Institute for Health Research (NIHR) Leeds Musculoskeletal Biomedical Research Unit,

Chapel Allerton Hospital,

Leeds,

UK

9. Centre for Rheumatology and Connective Tissue Diseases,

UCL Medical School,

Royal Free Campus,

London,

UK

10. Department of Rheumatology,

Freeman Hospital,

Newcastle upon Tyne Hospitals NHS Foundation Trust,

Newcastle upon Tyne,

UK

11. Department of Rheumatology,

Cambridge University Hospitals NHS Foundation Trust,

Cambridge,

UK

12. Department of Rheumatology,

Royal National Hospital for Rheumatic Diseases,

Royal United Hospitals NHS Foundation Trust,

Upper Borough Walls,

Bath,

UK

13. Department of Rheumatology,

Salford Royal NHS Foundation,

Salford,

UK

14. Department of Pharmacy and Pharmacology,

University of Bath,

Claverton Downs,

Bath,

UK

15. NIHR Manchester Musculoskeletal Biomedical Research Centre,

Central Manchester NHS Foundation Trust,

Manchester Academic Health Science Centre,

Manchester

UK

16. Photon Science Institute

University of Manchester,

Manchester

UK.

FUNDING

This study was funded by an Arthritis Research UK Clinical Studies Grant [20656] and an unrestricted educational grant from Actelion Pharmaceuticals Ltd. This study was supported

by Arthritis Research UK grant number (Centre for Epidemiology) 20380 and the Manchester Academic Health Sciences Centre (MAHSC).

COMPETING INTERESTS

CP Denton - GSK, Celgene, Actelion, Bayer, Sanofi, Roche-Genentech, Inventiva, 5,CSL Behring, GSK, Actelion, Roche-Genentech, Inventiva.

JD Pauling, has received unrestricted grant support (totalling £25000), speaker honoraria and consultancy fees from Actelion Pharmaceuticals.

AL Herrick has done consultancy work for Actelion, served on a Data Safety Monitoring Board for Apricus, received research funding and speaker's honoraria from Actelion, and speaker's honoraria from GSK.

AK Murray has received speaker honoraria from GSK and speaker honoraria and the grant associated with this study (referenced above) from Actelion.

All other authors declare no conflicts or competing interests.

Background: Objective and reliable outcome measures to facilitate clinical trials of novel treatments for systemic sclerosis (SSc)-related Raynaud's phenomenon (RP) are badly needed. Laser speckle contrast imaging (LSCI) and thermography are non-invasive measures of perfusion that show excellent potential. The purpose of this multi-centre study was to determine the reliability and validity of a hand cold challenge protocol using LSCI, standard thermography and low-cost mobile phone-based thermography.

Methods: Patients with RP secondary to SSc were recruited from 6 UK tertiary-SSc centres and underwent cold challenge on 2 consecutive days. Changes in cutaneous perfusion/temperature at each visit were imaged simultaneously using LSCI, standard and mobile phone thermography. Measurements included area under reperfusion/rewarming curve (AUC) and maximum perfusion/rewarming (MAX). Test-retest reliability was assessed using intraclass correlation coefficients (ICC). Estimated latent correlations assessed convergent validity of LSCI and thermography.

Results: 159 patients were recruited (84% female, 77% limited cutaneous, median age 63.3 years). LSCI and standard thermography both had substantial reliability, ICCs (95%CI) for AUC were 0.67(0.54-0.76) and 0.68(0.58-0.80) respectively, and for MAX were 0.64(0.52-0.75) and 0.72(0.64-0.81) respectively. Very high latent correlations (95% CI) were present for AUCs of LSCI and thermography [0.94(0.87-1.00)], and for AUCs of standard and mobile phone thermography [0.98(0.94-1.00)].

Conclusion: This is the first multi-centre study examining reliability and validity of cold challenge using LSCI and thermography in patients with SSc-related RP. LSCI and thermography demonstrated good potential as outcome measures. LSCI, standard and mobile phone thermography had very high convergent validity.

Systemic sclerosis (SSc)-related digital vasculopathy is painful and disabling, and has significant impact on quality of life. Raynaud's phenomenon (RP) occurs in most patients with SSc (96%) and is consistently the highest ranked symptom of SSc in terms of frequency and impact on daily function (1,2). In patients with SSc, RP often progresses to severe digital vasculopathy, with up to 50% of patients developing painful digital ulceration (3-11). Treatments are far from ideal and Cochrane and other reviews highlight the lack of evidence base for the treatment of both primary and SSc-related RP (12-15): one of the reasons for this is the lack of reliable outcome measures that are necessary to deliver successful clinical trials. Technological advances in laboratory measurements of blood flow (laser speckle contrast imaging [LSCI] and thermography [skin temperature- a pseudo measure of perfusion]) hold promise as objective measures (16,17). The Outcome Measures in Rheumatology (OMERACT) 6 report, describing the current status of outcome measure development for clinical trials in SSc, concluded that whether imaging techniques made the transition from research pathophysiology measurement techniques to outcome measures for RP was dependent upon 'whether data are published or available to show their validity' (18). The requirement for reliable outcome measures to facilitate highly powered clinical trials in SSc-related RP is now especially pertinent due to on-going novel drug developments (19-23). Whilst patient-reported outcome measures such as the Raynaud's condition score (RCS(24)) are well suited for later, phase III studies, objective non-invasive imaging techniques would provide confirmatory testing to inform stop-go decision-making in earlier phase II studies.

Our main aim was to determine whether LSCI and thermography, alongside a cold challenge of the hands, are sufficiently reliable and valid to allow their use as outcome measures in

multicentre clinical trials. Our primary objectives were to evaluate test-retest reliability and construct validity (25), which we defined as the ability of LSCI and thermography to measure important features of SSc-related digital vasculopathy. Our secondary objectives were to assess inter-observer reliability, and feasibility, of the techniques. Just prior to commencement of our study, mobile phone thermography came on the market as an imaging method, potentially offering a more cost-effective and portable alternative to LSCI and 'standard' thermography. Hence an additional secondary objective was to assess the utility of mobile phone thermography in comparison to standard thermography.

METHODS

Six UK tertiary-SSc centres took part in the study; individuals responsible for imaging and analysis attended a central training session prior to the start of recruitment; At least one person from each centre attended the training.

Patients

The study aimed to recruit 180 patients with SSc. Inclusion and exclusion criteria are listed in supplementary Table S1 and included current digital ulceration. The study was approved by Cambridgeshire and Hertfordshire NRES committee (15/EE/0083) and all patients gave written consent. Each visit took approximately one hour. All patients were recruited between 1/10/15 and 28/2/15 to minimise inter-individual variation related to season.

Imaging equipment

An LSCI (FLPI-2 [Moor Instruments, Axminster, UK] Figures 1a and 2b (16,17)) was leased to each centre. Five of the six centres used their own thermal cameras (referred to as 'standard thermography' Figure 1a, b and d (26)); a camera was leased to the sixth. A mobile phone/device-connectable thermography camera (FLIR One Figure 1a and 1c) and an Apple iPhone 5 was purchased for each centre along with all other cold challenge equipment (to minimise centre variation). To minimise differences between centres, equipment at each site was set-up according to strict guidelines for positioning to ensure images were taken in as similar manner as possible (angles/distances) and underwent a calibration protocol at the start and end of the study (carried out by a single person from the central site).. LSCI settings were adjusted for distance, frequency, duration, focus, intensity overlay, processing mode (high resolution), and colour image acquisition. Thermal cameras settings were adjusted for room temperature, distance to hands and skin emissivity. Mobile phone thermography settings were limited but set to matt.

Cold challenge

Patients were requested to wear light clothing and refrain from vigorous exercise, caffeine and alcohol for 4 hours prior to the assessment. Upon arrival patients were seated comfortably for 20 minutes and acclimatised; clinical research forms were completed. Immediately prior to the cold challenge, a baseline image of both hands (dorsal aspect) was taken with LSCI and both thermal cameras. As required for LSCI imaging, all images were acquired in low-lit rooms. The patient's hands were placed on a black, thermally insulated surface (one metre away from the thermal cameras and 70cm (+/- 5cm) from the LSCI). Small sticky dots were used to mark the location of each finger at baseline. Both hands were

nitrile gloved and immersed to the metacarpophalangeal joints for 1 minute into cooled water; 2 standard containers at 15 ± 1°C (measured by calibrated thermometer) one on either side of the patient. After the cold challenge, gloves were removed and hands returned to their original position on the insulating surface, secured by double sided sticky tape to avoid movement between images. Reperfusion/rewarming after cold challenge was imaged simultaneously by LSCI at 15 frames per minute and thermography (at 4 frames per minute) for 15 minutes (i.e contemporaneous measurement for 15 minutes post cooling). Mobile phone thermography did not allow for continuous video images to be obtained from which data could be extracted and thus single images were taken at set time points; baseline, 0 and 15 minutes after cold challenge. At the end of the 15 minutes one extra image was taken for LSCI and standard thermography to allow the gradient of the last data point to be calculated; thus a total of 225 images/scans were obtained for LSCI, 61 for thermography and 3 for mobile phone thermography during the 15 minutes of measurement. Analysis was performed in Moor Instruments Laser Perfusion Imager software version 4.0 for LSCI, and Research IR max version 4.2, [FLIR, Sweden] for standard and mobile thermography. Patients completed an RCS (0-10) at each visit ('RCS on the day'), measuring the severity and impact of their RP for that day (24).

The cold challenge was repeated one day later (Day 2) as close as possible to the same time of day to minimise variation due to circadian rhythms (27). The repetition over 2 *consecutive* days (i.e. approximately 24 hours) minimised any variations within individuals over time (e.g. menstrual cycle effects) and seasonal variation in weather (28). Five centres had one observer, one centre had 2. Each examiner re-examined same subject on days one and two.

(for example at the central site one observer imaged 60 patients, twice, on consecutive days) Figure 2 shows the study design.

Image analysis: summary measures of response

Image analysis was carried out locally by an internal non-blinded observer at each centre. These were the same observers that had taken the images. Regions of interest (ROIs, Figures 1b and 2b and c), were highlighted in the baseline (pre cold challenge) image and in sequential images for 15 minutes post cooling. The distal dorsal difference (DDD, measurement difference between dorsum and finger, (29,30) (with subscripted L, T or M for LCSI, thermography and mobile phone thermography respectively where applicable) was calculated for each finger at baseline. In the sequential images the ROIs were confined to the 8 distal phalanges. The area under the reperfusion/rewarming curve (AUC) for each finger was calculated manually, not by automation, (Figure 3 [standard thermography],3) from 61 post challenge images as well as the maximum blood flow/temperature after rewarming (MAX), and the gradient in the first 2 minutes (GRAD). Data were averaged for all fingers as in previous studies (16). For mobile phone thermography DDD was taken from the first of 3 images and AUC approximated by averaging over the latter two images. Analysis took less than one hour per participant, per visit.

Saved images and ROI local analysis data were also analysed by the blinded central observer. Mobile phone thermography image analysis was carried out only at the central site.

Feasibility

Feasibility was assessed at the end of the study by clinical scientist/technician opinion (ease of use and analysis 0-10 [0=difficult], preference of LCSI or thermography [standard or mobile] for acquiring and analysing images).

Room temperature measurement

A prerequisite of the cold challenge, and thus centre participation, was a temperature-controlled room at each centre. All measurements were taken in a temperature controlled room (aimed at 23 ± 2°C). Room temperatures were recorded with data monitors (TinyTag, Gemini Data Loggers, UK) to assess the impact of temperature on measurements, with an interest in examining whether reliability could be improved by achieving greater temperature control.

Edge effects from LCSI

It became apparent when the study began that the blood flow appeared to be lower at the edges of the LSCI images than in the centre. This implied that the distribution of the laser light across the hands was not equal, with less light incident towards the edges than at the centre of the image. If true then the consequence of this would be artificially lower value perfusion for little fingers (edge of image, Figure 4) compared to index fingers (centre of image) in the LSCI images. Thus this was investigated further as detailed below.

Statistical analysis

Analysis was performed using R version 3.2.3 (31).

Based upon a previous thermography study (16) 180 patients would allow us to estimate reliability to within 0.05. A full discussion of the sample size calculation and other aspects of the statistical analyses are provided in the supplementary material (extended statistical analysis).

Test-Retest Reliability of the techniques: Intraclass correlation coefficients (ICCs) were obtained using linear mixed effects models with each summary measure included as a dependent variable and with centre as a fixed effect.

Reliability between observers: The data over both visits for each patient were averaged and the resulting averages between central and centre-specific observer compared by calculating the difference and 95% CI for the paired means (Table S2). It is not possible to calculate a valid inter-observer ICC from this data, which would require at least some of the participants to have travelled to all sites for imaging and a large subset of images analysed by all observers (32).

Validity of the techniques: Convergent validity (one aspect of construct validity) was assessed using bivariate linear mixed models including fixed centre terms and separate random patient intercepts for 1) LSCI and standard thermography and 2) standard and mobile thermography. We estimated the latent correlation (which would be equal to one if

the techniques measured the same construct). For clarity, the mathematical representation of this joint model is provided in the supplementary material [statistical analysis (protocol)].

A post-hoc analysis was conducted where the responses to the RCS corresponding to the study day were related to the measurements, using linear mixed models.

Feasibility: Descriptive statistics were used to summarise the feasibility data.

Room temperature: The mean temperature for each patient visit was added to the mixed effects models for each summary measure. ICCs were recalculated, and compared with the previously calculated estimates.

Edge effects: This was investigated in a post-hoc analysis by calculating the trend across fingers for LSCI measurements and comparing these to thermography. Linear mixed models were used to assess any linear trends in the measurements from the index finger to the little finger. Fingers were numbered.Finger-level summary measures of response were then regressed on finger number for both LSCI and thermography; this linear approximation was crude but sufficient. Random intercept and slope terms were included to account for the fact that there was variation from patient to patient in these trends not attributable to the imaging techniques. Measurements were standardised prior to analysis, thereby allowing comparison to be made between LSCI and thermography.

RESULTS

159 patients were recruited (60 from the central centre, 16-20 from each of the others): 157 (99%) fulfilled the 2013 American College Rheumatology/European League Against Rheumatology criteria (34); median age 63.3 IQR (53.8-69.5) years; 123 (77%) lcSSc;

duration since first non-Raynaud's symptom 9.6 (4.5-17.4) years; 146 (93%) were on vasodilators (62 calcium channel blockers, 27 ACE inhibitors, 27 angiotensin II receptor antagonist, 22 Phosphodiesterase-5 inhibitor, 4 endothelin Receptor Antagonist, 1 Nitrates. 35 were on one or more vasodilator); 4 (3%) had previously had finger surgical debridements; 5 (3%) had previously had amputations; 30 (19%) had experienced ulcers in the last year.

Test-Retest Reliability of the techniques: There was at least moderate to substantial reliability for $DDD_{L,M,T}$ $AUC_{L,T,M}$ and $MAX_{L,T}$. $GRAD_{L,T}$ were fair to substantial (Table 1 (35)). A value of 0.7 could be considered high such that both MAX and AUC display strong convergence. Strength of reliability: 0.00 to 0.20 = slight; 0.21 to 0.40 = Fair; 0.41 to 0.60 = Moderate; 0.61 to 0.80 = Substantial; 0.81 to 1.00 = Almost perfect as per (33) although these classifications are to some extent arbitrary and should be treated as a rough guide.

~~Summary measures of responses to cold challenge obtained on days 1 and 2 at each centre for LCSI and thermography are shown in Table 2.~~

Reliability between observers: The data for each visit, observer and centre are displayed in Supplementary Table S2 and additionally at patient-level using ladder plots (Figure S1). If the measures were perfectly reliable the subplot for each centre would look like two identical ladders (but it is not expected that the plots will be identical between centres). Our data suggest that there were systematic differences between the central observer and centre 2 (and possibly centre 3) in extracting data from LSCI images. For thermography, agreement between the central and local observer was generally high for all centres, albeit with a large discrepancy for several patients for one of their visits.

Validity of the techniques: The latent correlation (95% CI) for LSCI and thermography (i.e. evidence that LCSI and standard thermography measure the same construct, in this case blood returning to the finger) was DDD: 0.65 (0.50 to 0.79); AUC: 0.94 (0.87 to 1.00) and MAX: 0.87 (0.77 to 0.95); but GRAD only 0.52 (0.33 to 0.70), Table 1. High latent correlation is indicative of convergent validity. A value of 0.7 could be considered high such that both MAX and AUC display strong convergent validity. Correlation between mobile thermography and standard thermography was also very high; 0.98 (0.94 to 1.00) for AUC and 0.90 (0.79 to 0.97) for DDD, Latent correlation between LSCI and FLIR was 0.86 (0.74 to 0.97) for AUC and 0.49 (0.29 to 0.66) for DDD. (Table 1).

With the exception of some weak evidence of decreasing DDD_T with increasing RCS (-0.15 DDD_T for a one point increase in RCS, on average, SE = 0.07) we found no evidence of correlation between the summary measures and RCS.

Feasibility: Standard thermography was deemed to be more feasible than LSCI, ~~In general, both techniques were deemed to be feasible (~~see discussion). The proportion of raters giving a score of 7 or above for ease of use (0=difficult–10=easy) was 50%for LSCI; 75%for standard thermography and 38%for mobile phone thermography. Ease of analysis was rated as 7 or above by 25% for LSCI and 50%for standard thermography. The number of centres preferring LSCI to thermography was 1 for acquiring and 1 for analysing images; preferring standard thermography 3 (acquiring) and 4 (analysing). The remaining centres showed no preference.

Room temperature: When included as a covariate, temperature was not associated with any of the summary measures as measured by either LSCI or thermography. Additionally, the ICCs were not affected by the inclusion of temperature in the analysis. This does not mean

that a regulated room temperature is not important but that small changes in temperature are acceptable (supplementary Table S3).

Edge effects: Moving from the thumb to the little finger, all of the AUC, MAX and GRAD trends were in the opposite direction for the two modalities, with a decrease for LSCI and an increase for thermography (supplementary Table S4). Both estimates for the DDD are positive, but this was attenuated for LSCI. This is consistent with an edge effect artificially producing lower values for the little fingers with LSCI. The cause of the edge effect was identified as the distribution of the light over the imaging area, due to LSCI being used at the upper limit of the suggested imaging distance in order to fit both hands into the imaging area. The data indicates that care must be taken to understand the variations over the field of view so that these can be accounted for; decreasing the field of view would minimise these results in future studies.

DISCUSSION

To date laser speckle imaging techniques and thermography have been insufficiently studied as outcome measures in clinical trials. Those studies in which they have been included show very little consistency in terms of protocol design (24, 36-38), choice of dynamic challenge and extracted outcome measures, making it difficult to compare results between studies or establish a standard protocol. The main finding of our study is that reliability of both LSCI and thermography were sufficiently high (AUC and MAX) for use as study outcome measures. The reliability of MAX_T was slightly superior to MAX_L. Other than this, there were no substantive differences in reliability between the two techniques.

AUC_M and DDD_M showed adequate reliability for use as outcome measures. Moreover, there was strong correlation between mobile phone and standard thermography data. The mobile phone thermography was added at a late stage in this project (since it had only just come on the market). Our reason for including it was primarily for feasibility assessment. While it is clear that further work is required to validate mobile thermography, the performance in the present study is highly encouraging because as a low-cost tool it could potentially be readily available for widespread use amongst rheumatologists.

Although not our primary objective, we examined differences between observers. Systematic differences between observers at different centres would not be particularly problematic for a multicentre randomised controlled trial (RCT), provided randomisation was stratified by centre. We note that this should be the default for any multicentre trial, since differences between centres may otherwise bias the estimated treatment effect. This is particularly true in small populations, since simple randomisation is less likely to produce balance within centres. Standardised training would reduce measurement variation across centres and centralised blinded extraction and analysis of LSCI data might also minimise variation by removing multi-observer differences in an RCT setting. Given the small sample size at each centre we are unable to determine whether truly systematic differences were observed. Ideally, a study to assess inter-observer reliability would involve participants having images analysed by all observers.

Convergence between the techniques was shown to be very high for AUC and MAX (particularly for AUC). This provides evidence that the same underlying construct is being measured when using these summaries of response. Convergence appeared to be weaker

(although still moderate) for DDD. Convergence was weakest for GRAD which may reflect a lag between tissue re-perfusion (measured with LSCI) being translated into skin re-warming (measured using thermography) during the 2 minutes immediately following cold challenge.

Since there is no gold standard to compare either imaging technique to, and we are comparing two techniques that measure perfusion by very different methods (skin temperature and a measure of red blood cell concentration and speed by light) it is possible to measure convergence between these techniques for validity (25). It would be unlikely for these two techniques to converge whilst also being poor measures since they would both have to be deficient in distinct but very specific ways so as to bring the erroneous observations into alignment. Therefore we can conclude in this instance that their convergence implies validity.

The OMERACT review of 2003 (18) assessed the validity of several non-invasive techniques as possible objective outcome measures but none was deemed ready for use in clinical trials. These included nailfold capillaroscopy, a well-established diagnostic technique is now included in the diagnostic criteria (34) to differentiate primary and secondary Raynaud's phenomenon. The microscopy technique allows visualisation of cutaneous capillaries at the nailbed and identification of structural change characteristic of SS. This is not a substitute though for functional measures of flow (although functional flow and oxygenation have recently been reported). Plethysmography allows the change in vascular volume to be measured (i.e. detection of a pulse) in combination with cold challenge. The technique can measure full fields in the same ways as laser speckle but remains unvalidated. There was no relationship between the summary measures and the RCS on the day of the study visits, for

either LSCI or thermography. Patient-centred outcome measures are crucial for evaluating the effectiveness (rather than just the efficacy) of treatments. However, patient-centred outcomes often comprise more 'noise' compared to more objective measures of response, and therefore necessitate larger sample sizes to ensure adequate power in clinical trials. For small populations, there is therefore a tension between direct relevance to patients and feasibility of conducting a trial. One solution may be to power studies on the basis of objective measures such as those considered here, and to additionally (and consistently) report patient-centred outcomes to facilitate an eventual meta-analysis. Another might be to seek confirmatory evidence for the vasodilatory potential of candidate interventions using objective measures before proceeding to larger phase 3 clinical trials

The relationship between two measures is limited by the reliability of each (39). While relative stability of RCS has been observed between baseline and follow-up in clinical trials/studies (38,40), there has been little work formally assessing its intra-individual reliability.

Regarding feasibility, comments were made regarding LSCI and its sensitivity to movement, vibrations and lighting, indicating the importance of environmental conditions. For the mobile phone thermography, present limitations include battery life (LSCI and thermography were mains/long-life battery powered), fixed focussing distance and lack of analysis for video images, as well as mounting difficulties; however, if the mobile and standard thermography correlation can be replicated in future studies, these limitations may be acceptable in light of the lower cost and ambulatory (convenient) nature of the technique. When comparing feasibility of LSCI versus thermography, it should be noted that most centres were familiar with thermography but not LCSI, and that this may have influenced assessment of feasibility.

One limitation of the study was that we did not recruit the planned number of participants, due to a seventh centre not participating as planned. However, the study was designed to be robust to under-recruitment. Although the 95% CIs for our estimates are wider than they would have been had the target been met, we were still able to demonstrate good enough reliability and convergent validity of AUC and MAX, to observe differences indicative that that the performance of DDD was weaker, and to show that the performance of GRAD was relatively poor.

In conclusion, our design was relatively pragmatic, with the aim of establishing the performance of the different techniques as they would be employed in a multicentre clinical trial. Our study successfully established a working group of tertiary-SSc centres, and together the group developed a consensus calibration and cold challenge protocol. The summary measures AUC and MAX both displayed good reliability and strong convergent validity. There was a possible advantage of thermography in relation to the reliability of MAX, although this was not definitive. We found evidence of edge effects when using LSCI although our summary measures appeared to be quite robust to these in relation to reliability, perhaps suggesting that these effects were fairly consistent (methods and results discussed in supplementary data). The study has also confirmed that small variations in room temperature are acceptable and that, subject to further validation, mobile phone cameras may be a suitable, affordable, highly portable alternative to more expensive standard imaging equipment (although mobile phones are battery operated and with less functionality [at present] than larger thermal cameras). The mobile phone data obtained in

this study will facilitate the design of future validation studies into mobile phone thermography-derived outcome measures. Although the design precluded formal assessment of inter-observer reliability there was a suggestion of systematic differences between the central observer and observers at some centres, highlighting the importance of image analysis training and potentially a role for centralised or automated image analysis. For multicentre RCTs, we would also recommend that where possible/appropriate, randomisation be stratified by centre to balance any centre effects and prevent bias.

In summary LSCI and thermography should now be incorporated as secondary outcomes in upcoming treatment efficacy trials. This will allow an assessment of responsiveness to treatment as well as longitudinal validity. The present study leads us to recommend the summary measures AUC and MAX, measured using both thermography and LSCI (but especially using thermography), as suitable outcome measures for RCTs in SSc-related RP.

ACKNOWLEDGEMENTS

This study was funded by an Arthritis Research UK Clinical Studies Grant [20656] and an unrestricted educational grant from Actelion Pharmaceuticals Ltd. This study was supported by Arthritis Research UK grant number (Centre for Epidemiology) 20380 and the Manchester Academic Health Sciences Centre (MAHSC). The authors would like to acknowledge the assistance given by IT Services and the use of the Computational Shared Facility at The University of Manchester.

We are grateful to the trial steering and data monitoring committee members: Dr Mohammad Akil, Prof David D'Cruz and the late Prof Peter Wells.

We also thank Dipa Ghedia at the London site and Sook Eng at the Leeds site for patient recruitment, Darren Hart at the Bath site for imaging and analysis and Anita Furlong and Tracey Drayton at the Cambridge site for imaging.

We wish to thank the UK Scleroderma Study Group for their advice and support in the development and running of this study and Moor Instruments and Thermal Vision Research for their advice and training.

REFERENCES

1. Walker UA, Tyndall A, Czirjak L, Denton C, Farge-Bancel D, Kowal-Bielecka O, et al. Clinical risk assessment of organ manifestations in systemic sclerosis: a report from the EULAR Scleroderma Trials and Research group database. Ann Rheum Dis 2007;66:754-63.
2. Willems LM, Kwakkenbos L, Leite CC, Thombs BD, van den Hoogen FH, Maia AC et al. Frequency and impact of disease symptoms experienced by patients with systemic sclerosis from five European countries. Clin Exp Rheumatol 2014;32:S-88-93
3. Rodnan GP, Myerowitz RL, Justh GO. Morphological changes in the digital arteries of patients with progressive systemic sclerosis (scleroderma) and Raynaud's phenomenon. Medicine 1980;59:393–408.
4. Herrick A. Diagnosis and management of scleroderma peripheral vascular disease. Rheum Dis Clin North Am 2008;34:89-114.
5. Steen VD, Powell DL, Medsger TA. Clinical correlations and prognosis based on serum autoantibodies in patients with systemic sclerosis. Arthritis Rheum 1988;31:196-203.
6. Della Rossa A, Valentini G, Bombardieri S, Bencivelli W, Silman AJ, D'Angelo S, et al. European multicentre study to define disease activity criteria for systemic sclerosis. Clinical and epidemiological features of 290 patients from 19 centres. Ann Rheum Dis 2001;60:585-91.
7. Ferri C, Valentini G, Cozzi F, Sebastiani M, Michelassi C, La Montagna G, et al. Systemic sclerosis: demographic, clinical and serologic features and survival in 1,012 Italian patients. Rheumatol 2002;81:139-53.
8. Mawdsley AH. Patient perception of UK scleroderma services – results of an anonymous questionnaire. Rheumatol 2006;45:1573.

9. Tiev KP, Diot E, Clerson P, Dupuis-Siméon F, Hachulla E, Hatron PY, et al. Clinical features of scleroderma patients with or without prior or current ischemic digital ulcers: post-hoc analysis of a nationwide multicenter cohort (ItinerAIR-Sclerodermis) J Rheum 2009;36:1470-6.
10. Khimdas S, Harding S, Bonner A, Zummer B, Baron M, Pope J; et al. Associations with digital ulcers in a large cohort of systemic sclerosis: results from the Canadian Scleroderma Research Group Registry. Arth Care Res 2011:63;142-9.
11. Ennis H, Vail A, Wragg E, Taylor A, Moore T, Murray A, et al. A prospective study of systemic sclerosis-related digital ulcers: prevalence, location, and functional impact. Scand J Rheumatol. 2013;42:483-6.
12. Ennis H, Hughes M, Anderson ME, Wilkinson J, Herrick AL. Calcium channel blockers for primary Raynaud's phenomenon. Cochrane Database of Systematic Reviews 2016. DOI: 10.1002/14651858.CD002069.pub5.
13. Herrick AL. Raynaud's phenomenon (secondary). BMJ Clinical Evidence (Online) 2008; 1125.
14. Stewart M, Morling JR. Oral vasodilators for primary Raynaud's phenomenon. Cochrane Database of Systematic Reviews 2012, Issue 7. Art. No.: CD006687.
15. Garcia de la Pena Lefebvre P, Nishishinya MB, Pereda CA, Loza E, Sifuentes Giraldo WA, Román Ivorra JA, et al. Efficacy of Raynaud's phenomenon and digital ulcer pharmacological treatment in systemic sclerosis patients: a systematic literature review. Rheumatol Int 2015;35:1447-1459.
16. Murray AK, Moore TL, Manning JB, Taylor C, Griffiths CE, Herrick AL. Noninvasive imaging techniques in the assessment of scleroderma spectrum disorders. Arthritis Rheum 2009;61:1103-11.

17. Pauling JD, Shipley JA, Raper S, Watson ML, Ward SG, Harris ND, et al. Comparison of infrared thermography and laser speckle contrast imaging for the dynamic assessment of digital microvascular function. Microvasc Res 2012;83:162-7.
18. Merkel PA, Clements PJ, Reveille JD, Suarez-Almazor ME, Valentini G, Furst DE, et al. Current status of outcome measure development for clinical trials in systemic sclerosis. Report from OMERACT 6. J Rheumatol 2003;30:1630-47.
19. Herrick AL. Secondary Raynaud's phenomenon. BMJ Clinical Evidence 2008;09:1125.
20. Cerinic M, Denton CP, Furst DE, Mayes MD, Hsu VM, Carpentier P, et al. Bosentan treatment of digital ulcers related to systemic sclerosis: results from the RAPIDS-2 randomised, double-blind, placebo-controlled trial. Ann Rheum Dis 2011;70:32-8.
21. Fava A, Wung PK, Wigley FM, Hummers LK, Daya NR, Ghazarian SR, et al. Efficacy of Rho kinase inhibitor fasudil in secondary Raynaud's phenomenon. Arthritis Care Res 2012;64:925-9.
22. Khanna D, Denton CP, Merkel PA, Krieg T, Le Brun FO, Marr A, et al. Effect of Macitentan on the Development of New Ischemic Digital Ulcers in Patients With Systemic Sclerosis: DUAL-1 and DUAL-2 Randomized Clinical Trials. JAMA. 2016;315:1975-88.
23. Seibold JR, Wigley FM, Schiopu E, Denton CP, Silver RM, Steen VD, et al. Digital ulcers in SSc treated with oral treprostinil: a randomized, double-blind, placebo-controlled study with open-label follow-up. J Scleroderma Relat Disord 2017; 2: 42-9.
24. Merkel PA, Herlyn K, Martin RW, Anderson JJ, Mayes MD, Bell P, et al. Measuring disease activity and functional status in patients with scleroderma and Raynaud's phenomenon. Arthritis Rheum 2002;46:2410-20.

25. Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use. Oxford University Press, USA; 2014 Oct 30.

26. Clark S, Dunn G, Moore T, Jayson M 4th, King TA, Herrick AL. Comparison of thermography and laser Doppler imaging in the assessment of Raynaud's phenomenon. Microvasc Res 2003;66:73–6.

27. Houben AJ, Slaaf DW, Huvers FC, de Leeuw PW, Nieuwenhuijzen Kruseman AC, Schaper NC. Diurnal variations in total forearm and skin microcirculatory blood flow in man. Scand J Clin Lab Invest 1994;54:161-8.

28. Bartelink ML, Wollersheim H, Theeuwes A, van Duren D, Thien T. Changes in skin blood flow during the menstrual cycle: the influence of the menstrual cycle on the peripheral circulation in healthy female volunteers. Clin Sci (Lond) 1990;78:527-32.

29. Clark S, Hollis S, Campbell F, Moore T, Jayson M, Herrick A. The "distal-dorsal difference" as a possible predictor of secondary Raynaud's phenomenon. J Rheumatol 1999;26:1125-8.

30. Anderson M, Moore T, Lunt M, Herrick AL. The 'distal-dorsal difference': a thermographic parameter by which to differentiate between primary and secondary Raynaud's phenomenon. Rheumatol 2007;46:533-8.

31. R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

32. Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Physical therapy 1994 Aug 1;74(8):777-88.

33. LeRoy EC, Black C, Fleischmajer R, Jablonska S, Krieg T, Medsger TA Jr, et al. Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. J Rheumatol 1988;15:202–5.

34. van den Hoogen F, Khanna D, Fransen J, Johnson SR, Baron M, Tyndall A, et al. 2013 classification criteria for systemic sclerosis: an American College of Rheumatology/European League against Rheumatism collaborative initiative. Arthritis Rheum 2013 Nov;65(11):2737-47.

35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-174.

36. Pauling J, Shipley J, Harris N, McHugh NJ. Use of infrared thermography as an endpoint in therapeutic trials of Raynaud's phenomenon and systemic sclerosis. Clin Exp Rheumatol 2012;30:S103-15.

37. Allen J, Howell K. Microvascular imaging: techniques and opportunities for clinical physiological measurements. Physiol Meas 2014;35:R91-R141.

38. Pauling JD, Shipley JA, Hart DJ, McGrogan A, McHugh NJ. Use of Laser Speckle Contrast Imaging to Assess Digital Microvascular Function in Primary Raynaud Phenomenon and Systemic Sclerosis: A Comparison Using the Raynaud Condition Score Diary. J Rheumatol 2015;42:1163-8.

39. Fleiss JL. Reliability of Measurement. Ch.1 in Design and Analysis of Clinical Experiments. Volume 73 of Wiley Classics Library. John Wiley & Sons, 2011.

40. Gladue H, Maranian P, Paulus HE, Khanna D. Evaluation of test characteristics for outcome measures used in Raynaud's phenomenon clinical trials. Arthritis Care Res 2013;65:630-6.